ACCT420 Forecasting & Forensic Analytics

AY2018-19 Term 1

Project Topic: Google Analytics: Customer Revenue Prediction

Title: An Amateur's Guide to Data Analysis

Prepared by: Team 2

Gavin Ong Ze Kai (S9445712B)

Jerome Jeevan Naidu (S9541681J)

Ng Yuan Cheng (S9611095B)

Then Jia Hui (S9642532E)

Yeo Pei Si Clarice (S9545522J)

Date of submission: November 2018

For:

Professor Richard M. Crowley

Table of Contents

1. Executive Summary

This report documents our group's participation in an ongoing Kaggle competition to analyse a Google Merchandise Store customer dataset to predict the revenue per customer.

We started off by conducting an Exploratory Data Analysis (EDA) on the train data to eliminate variables with large amounts of missing data and to analyse the usefulness of variables we expect to have an impact on transaction revenue.

Running LASSO on the variables we found to be useful gave us a root mean square error (RMSE) of 1.6410, placing us at 2928th in the competition. However, our group realised that a regression analysis may not be the best option as some of the variables are categorical variables which LASSO is not equipped to handle. Understanding that this is in fact a classification and regression data science project, we then opted to use the Random Forest algorithm. This gave us a RMSE of 1.6472. Using the %IncMSE measure in Random Forest also helped us to identify variables that higher predictive power - pageviews, subContinent and hits.

Reflecting on the results of our first two models, we realised that due to the very small proportion of visits with any revenue (1.27%), a two-part model that first identifies whether the customer will make a transaction before predicting the revenue of such a customer will be more accurate.

Our group proceeded to do a second EDA by splitting the train data into visitors who made transactions and visitors who did not. We discovered a few variables that have predictive power of transaction revenue and hence feature engineered 2 variables (browser == Chrome, and deviceCategory == desktop & operatingSystem == "Chrome OS", "Macintosh" or "Linux"). Using this revised Random Forest Model, RMSE was reduced to 1.5616.

We decided to use Gradient Boosting Machines (GBM) in our final model as it performs better than Random Forest if the parameters are carefully tuned. Using the same variables, we achieved an ultimate Kaggle RMSE of 1.5387, ranking us at 2739th.

2. Data Cleaning

Before analysis of the data, we performed the following steps to clean the dataset provided:

1. Flatten columns[1]
2. Change all observations with NA-related strings to NA[2]
3. Assign 0 to the NA values of transaction revenue (for train data set), pageviews, bounces and newVisits as these are numerical variables that should either be zero or have a number
4. Assign “Not Set” to the NA values of other categorical variables
5. Convert appropriate variables to date, factor and numeric
6. Apply log transformation on transactionRevenue from train data set

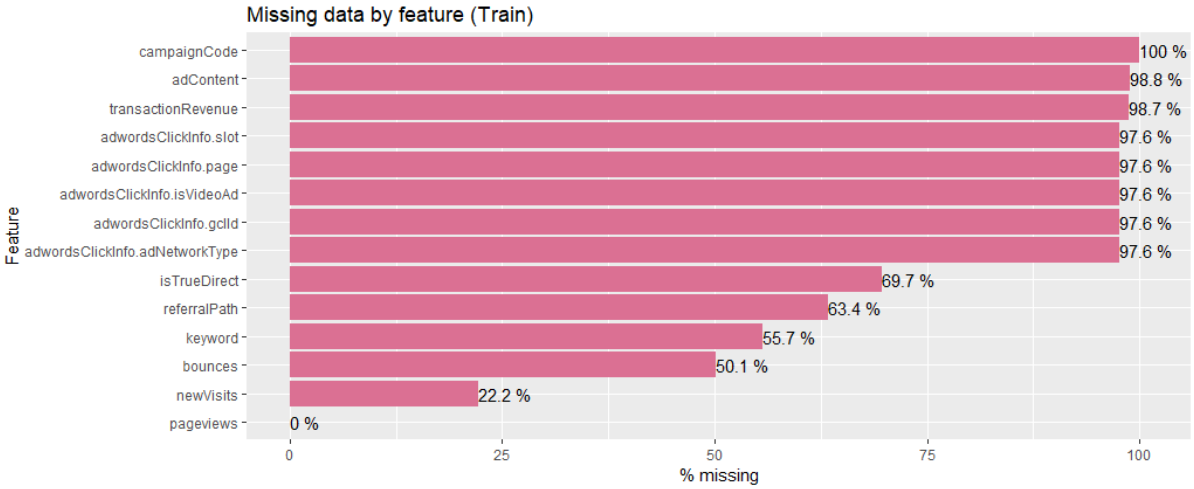
3. Exploratory Data Analysis

We will be conducting exploratory data analysis using the train data. This would help us to gain a better understanding of our data and to choose the variables to be included in our model later on. It also helps us to understand the missing values or discrepancies we have in our data so that we can do some data cleaning before we proceed on with modelling.

[1] Code taken from: https://www.kaggle.com/mrknoot/gstore-crafting-models-manual-features-xgb
[2] Code taken from: https://www.kaggle.com/mrknoot/gstore-crafting-models-manual-features-xgb

3.1 Missing values

From the train data, there are 14 variables that have missing values as seen from the graph below.



We will be removing campaignCode, all adwordsClickInfo.(Parameters), adContent, due to the large amount of missing values.

These following variables will be removed as well due to them having only category == “not available in demo dataset” : browserVersion, browserSize, operatingSystemVersion, mobileDeviceBranding, mobileDeviceModel, mobileInputSelector, mobileDeviceInfo, flashVersion, language, screenColors, screenResolution. While for socialEngagementType, the only level is not socially engaged and hence it will be removed as well.

The variable “visits” only have one value which 1 hence we will be removing it as well.

We will be removing these variables - region, Metro, City, cityId, latitude, longitude, networkLocation, because continents and subcontinents variables will be used as location information variables. As these variables have a large number of levels, the number of observations for each level may not be sufficient and hence negatively impacting the usefulness of these variables in predicting revenue for each sessionId.

3.2 Duplicates in sessionId

The total number of observations in train set is 903,653. Our group checked for the number of unique sessionId since it is supposed to be an unique identifier. The output returned is 902,755 which means there are 898 duplicates in sessionId.

All duplicated sessionId appeared twice. However, when we checked through the duplicated observations, they are almost the same except for visitStartTime, visits, hits, pageviews, bounces, newVisits. A possible explanation is that the users might have opened another browser. It could be due to time zone differences where it is a single session but the session extends beyond midnight to the next day.

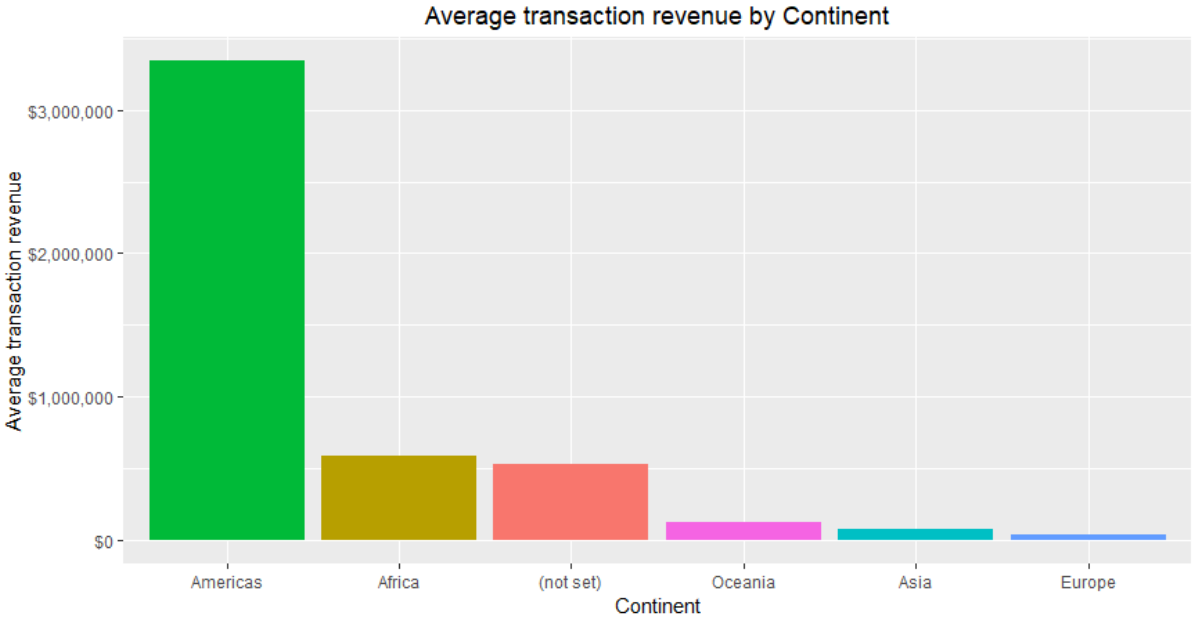
3.3 Plotting of Variables against Average Transaction Revenue

We are taking the average of transaction revenue as our y-axis for the following plots in this section. This is because the number of transactions with revenue takes up only 1.27% (11,515 out of the 903,653 observations) of the train dataset. If we were to just analyse by the revenue earned, it may result in an inaccurate representation as a single large spending by a certain sessionId will cause the analysis to be skewed. On the other hand, if the number of users is used as the dependent variable, we will only be analysing the number of users but not the amount spent each time.

Hence, the average transaction revenue is a better indicator to use when comparing the performance of each independent variables and we will use it to further predict the revenue earned for each sessionId.

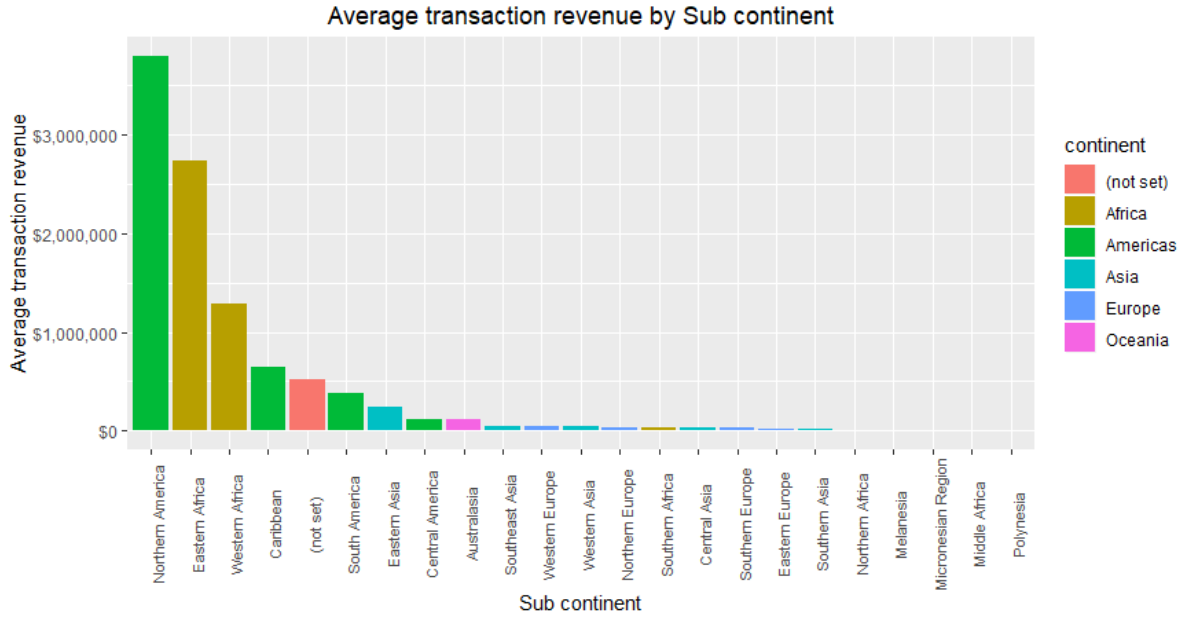
The following graphs that were plotted are based on a pick of variables which we expect to be correlated with transaction revenue.

3.3.1 Continents



From the bar graph, the Americas have the highest average transaction revenue while Europe has the lowest. The other variables' average transaction revenue only takes up less than a sixth of the average revenue transacted from the Americas continent. It is possible that visitors on the Google Store from these continents are not drawn to Google products, hence the amount transacted is rather low.

3.3.2 Sub Continents

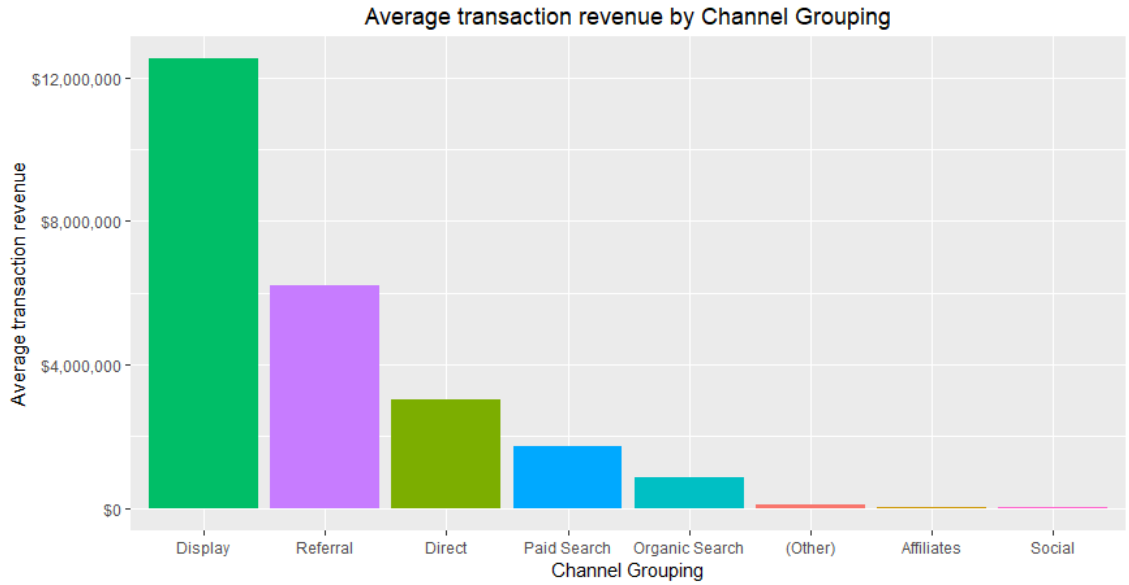


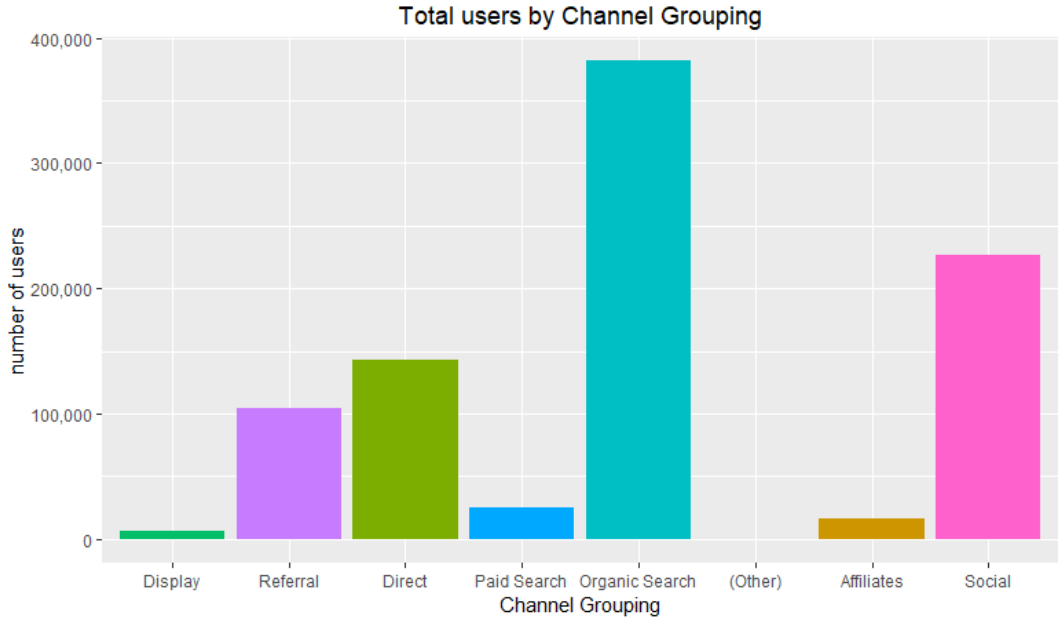
To gain a better understanding from the previous graph, we analysed the average transaction revenue by subcontinent instead. This breaks the data further into smaller groups. From *3.3.1 Continents*, we concluded that the Americas took the top spot in having the highest average transaction revenue. However, as we split the data further into subcontinents, Northern America actually has the highest average transaction revenue, followed by Eastern Africa and Western Africa.

However, from the graph we can infer that the other parts of the America continents, such as Caribbean, South America, Central America, do not generate a significant amount of revenue. Hence, subcontinent may be a more specific and accurate indicator of revenue transacted as compared to continent. This is because when continents are used, we may be quick to conclude that all of Americas' subcontinents are significant contributors of the revenue earned.

3.3.3 Channel Grouping

Channels are the methods how a user came to the website; while channel groupings are categories of these traffic sources. The definitions collated below in the Table 1 are given by Google Analytics. Most of these channel groupings are defined by Medium, Source, Social Source Referral or Ad Distribution Network. According to Google Analytics (2018), Source refers to the origin of the traffic such as a search engine or referral website. Medium refers to nature of the traffic such as "organic search", "cost per click" (cpc), "cost per mile" (cpm), or "none" (for direct traffic).[3]





[3] Google Analytics. (2018). Analytics Help - Traffic Source Dimensions. Retrieved from https://support.google.com/analytics/answer/1033173?hl=en&ref_topic=6010089

Channel	Description
Display	Medium matches regex ^(display|cpm|banner)$ OR Ad Distribution Network exactly matches Content
Referral	Medium exactly matches referral
Direct	Source exactly matches direct AND Medium exactly matches (not set) OR (none)
Paid Search	Medium matches regex ^(cpc|ppc|paidsearch)$ AND Ad Distribution Network does not exactly match Content
Organic Search	Medium exactly matches organic
Affiliates	Medium exactly matches affiliate
Social	Social Source Referral exactly matches Yes OR Medium matches regex ^(social|social-network|social-media|sm|social network|social media)$

Table 1. Channel Definitions *(from Google Analytics)*[4]

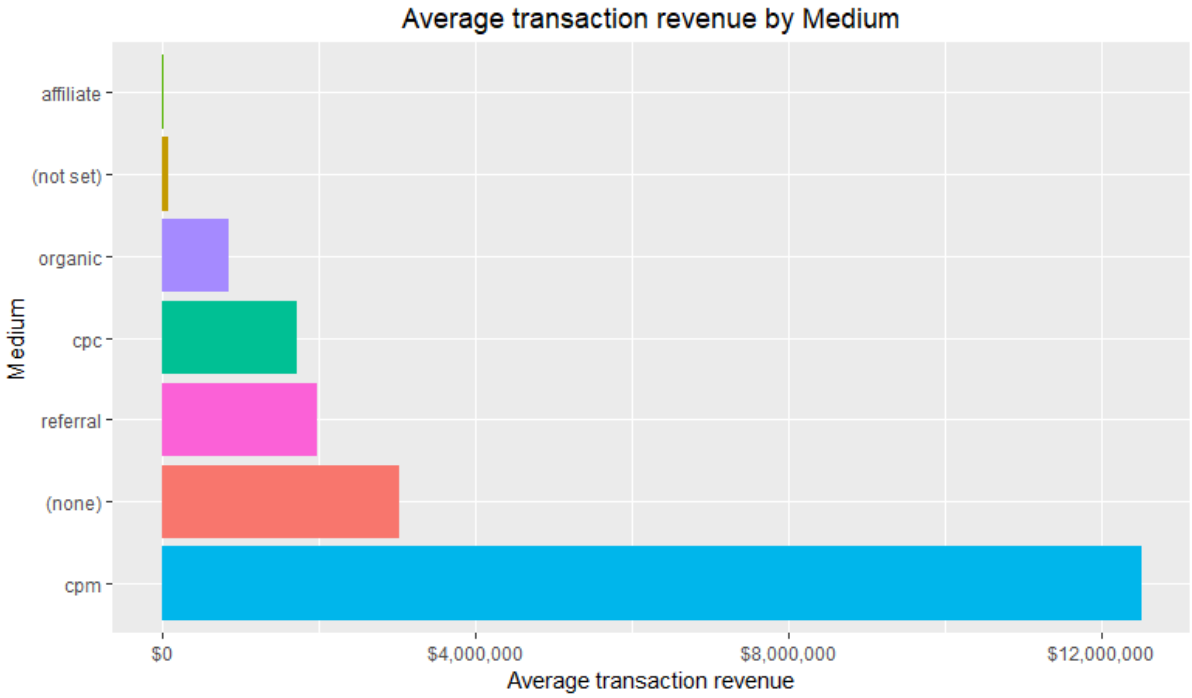
From the plots, 'Display' has the highest average transaction revenue but the number of users that accessed via 'Display' is low. This is possibly due to the advertisements shown are targeted specifically according to the user's search history or preferences. Hence, increasing the chances of users buying the products on the store.

Affiliate marketing have helped increased the number of users, however, the average transaction revenue earned was very low. Hence, this marketing method may seem like an ineffective. On the other hand, paid search is a more effective advertising campaign as the number of visits and average transaction revenue were higher.

Nevertheless, most of the users visited the website from Organic Search, little revenue was generated from this group of people. This may be a result of users wanting to just browse through the products without the intention of buying.

[4] Google Analytics. (2018). Analytics Help - Default channel definitions. Google. Retrieved from https://support.google.com/analytics/answer/3297892

3.3.4 Medium Dimension

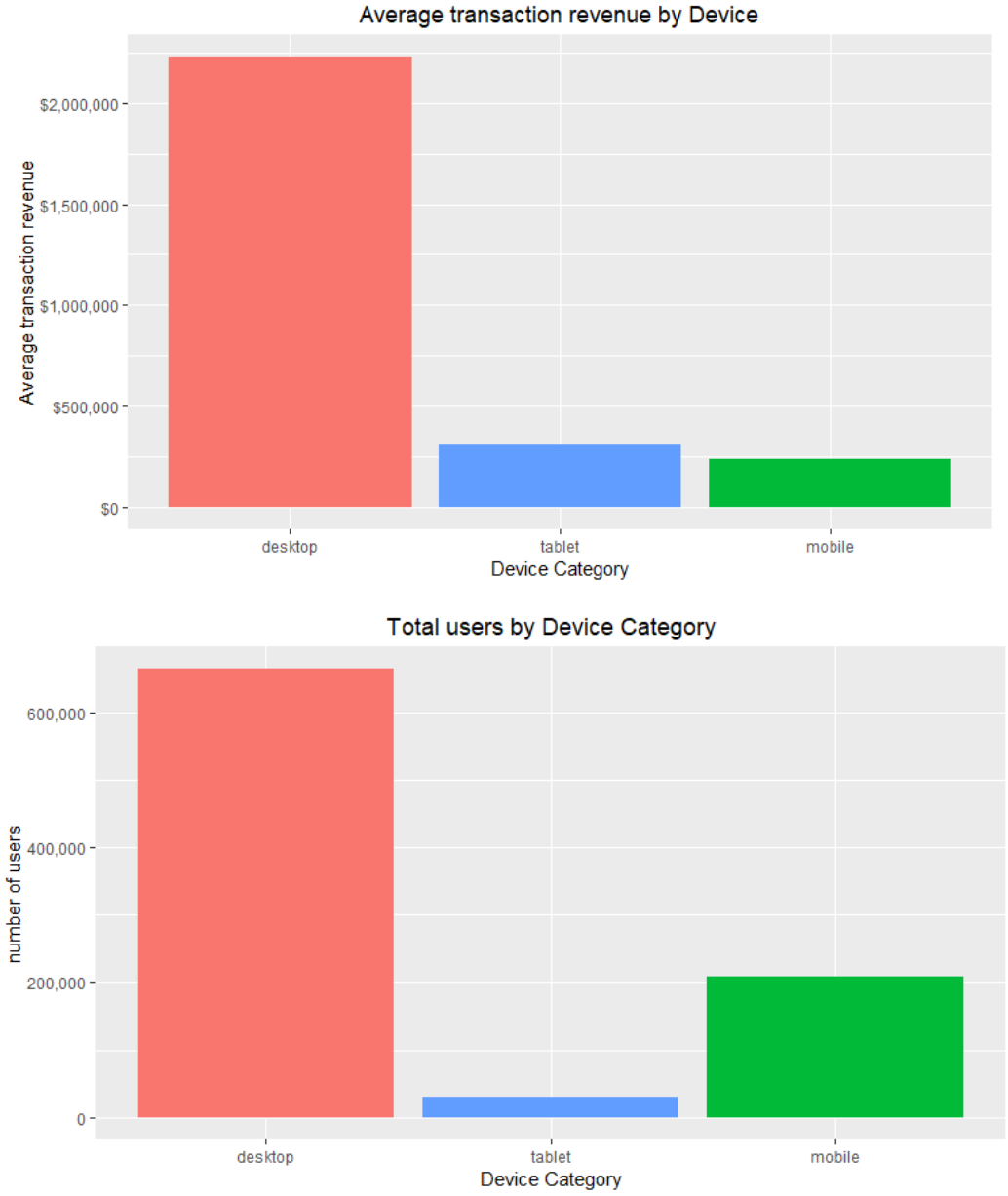


The medium dimension is closely linked to the Channel Groupings variable as defined by Google Analytics and medium can determine the type of channel grouping.

From the plot, cpm generates the highest average transaction revenue which matches the result we obtained from the *3.3.3 Channel Grouping plot.* Cpm is a type of digital marketing and it refers to the cost of 1,000 advertisement impressions on one webpage.[5] Since cpm is grouped under the Display channel, we can infer that cpm is the largest variable under the 'Display' channel and the biggest contributor of the average transaction revenue. Therefore, cpm was one of the most effective digital marketing strategy when compared to the other variables.

[5] Investopedia. (n.d.) Cost per Thousand - CPM. Retrieved from https://www.investopedia.com/terms/c/cpm.asp

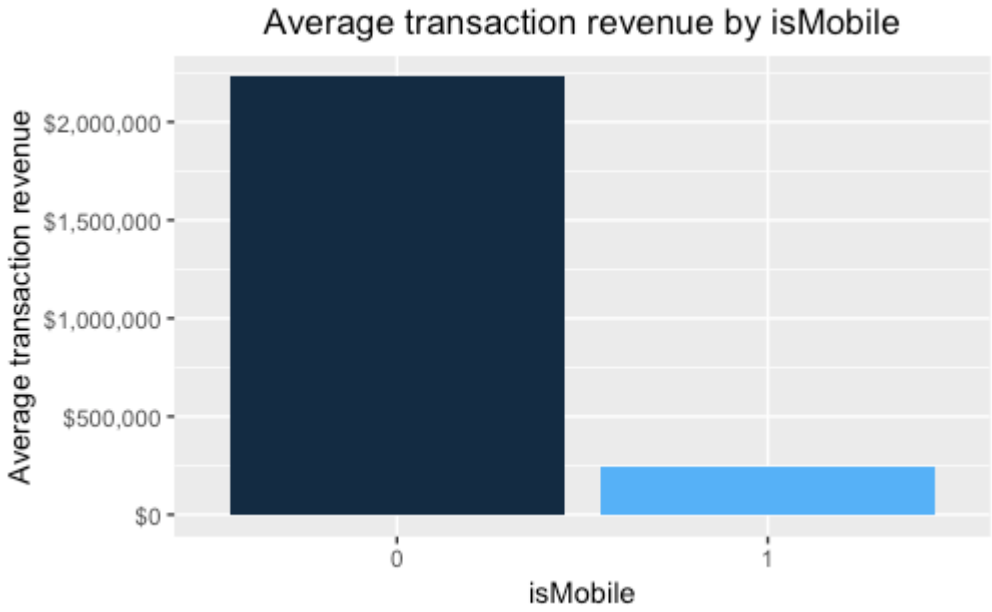
3.3.5 Device Categories



From this plot, we derived that users using desktop generates the highest average transaction revenue. Perhaps this is because it is easier and more common for users to use desktops to make an online transaction as compared to the other two devices. According to an article on Forbes, people tend to do deep online research on a desktop instead of a mobile device.[6] Furthermore, people do not usually made their purchases on a mobile device and it is merely used as a research tool throughout the purchase process. Hence, there is a higher transaction revenue for desktop user.

[6] Forbes Communication Council. (2018, Mar). 12 Major Differences Between Mobile And Desktop Marketing. Forbes. Retrieved from https://www.forbes.com/sites/forbescommunicationscouncil/2018/03/23/12-major-differences-between-mobile-and-desktop-marketing/#76ccd7eb5d6b

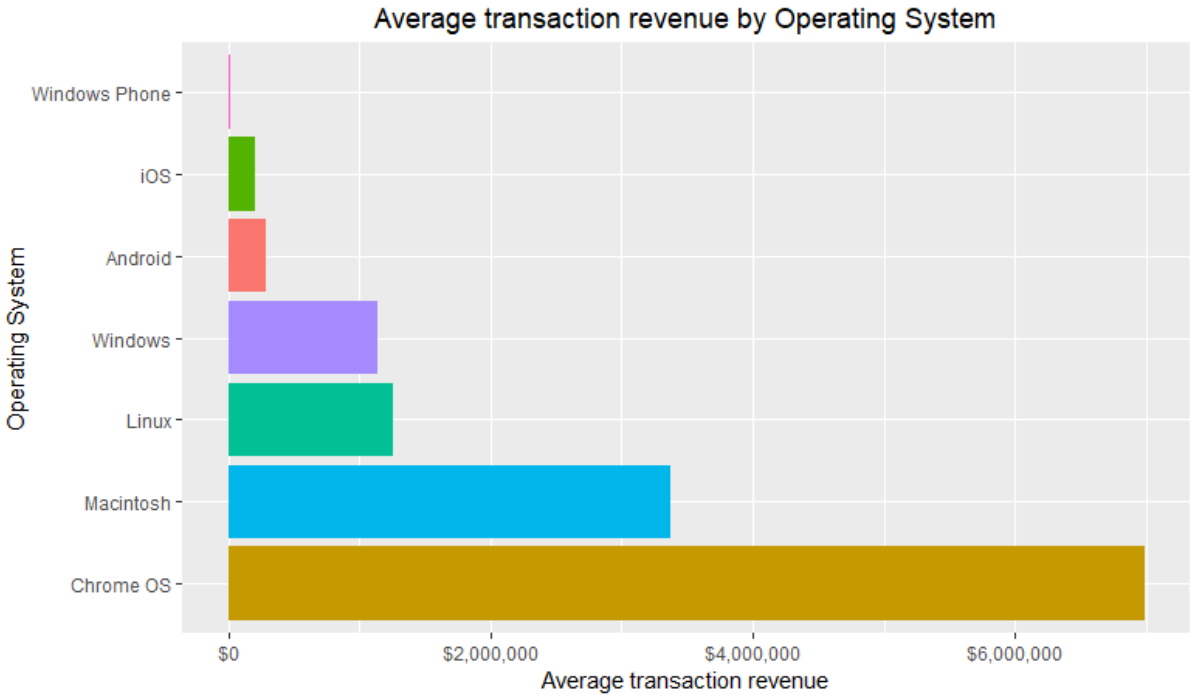
3.3.6 isMobile



isMobile is highly correlated to *3.3.5 Device Categories*. 0 refers to average transaction revenue derived from customers using a non-mobile device (desktop). On the other hand, 1 refers to the average transaction revenue generated from customers using a mobile device, which includes mobile and tablet.

Higher average transaction revenue was generated when non-mobile devices were used and this result corresponds to the output in *3.3.5 Device Categories*.

3.3.7 Operating Systems



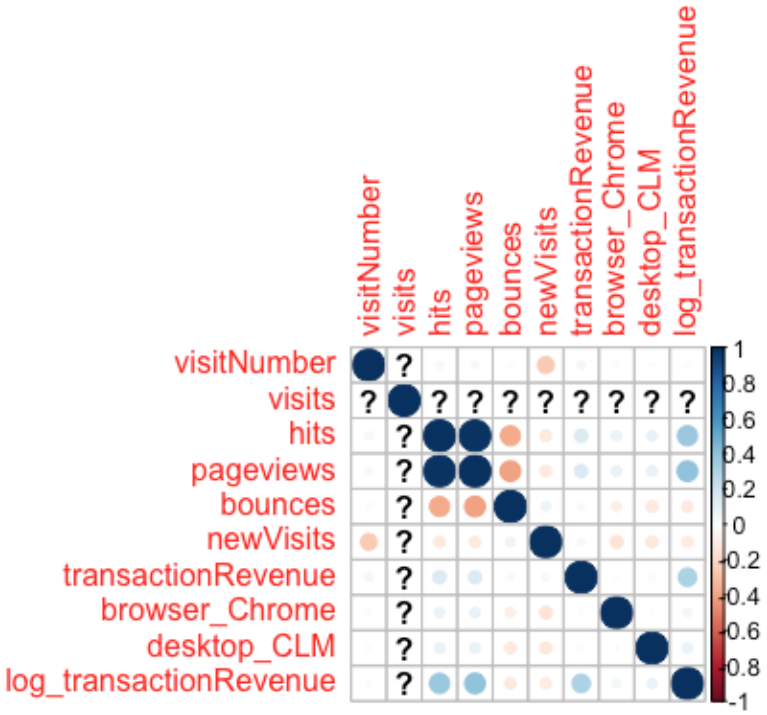
This graph shows the 7 Operating Systems that have the highest average transaction revenue. Chrome OS takes the lead, followed by Macintosh and Linux in average transaction revenue.

Chrome OS is an operating system that was developed by Google. Hence, users who prefer Google's product might have spent more and this led to a higher average transaction revenue amount.

3.3.8 Hits and Pageviews

With higher hits and pageviews, the chances and amount of revenue transacted will increase. This maybe due to the fact that people were interested in the merchandises and they were spending more time on the website, looking for products to buy. Hence, we will be including these variables in our models.

3.4 Correlation Matrix



As shown on this correlation matrix, only hits and pageviews are highly correlated. We should perhaps (1) decide whether we should use these two correlated variables in our model because it might add unnecessary weight or (2) use an algorithm that takes care of correlated variables.

The question mark that is seen for 'visits' is due to a constant value of '1' with no standard deviation, which makes sense since each observation is counted as '1' visit and not a sum of visits for that particular user. Thus, corrplot is not able to plot the correlation relating to 'visits'.

4. Model Selection

4.1 Choice of error metric

We have opted to use root mean squared error as our error metric across our models because the Kaggle competition is using it. It is also common to use RMSE for regression-type models.

4.2 LASSO

Based on the variables identified above, our group set out to construct a regression model to predict the transaction revenue. Our group felt that there was a distinct positive relationship between our dependent variable (transaction revenue) and our independent variables (pageviews, hits, isMobile, channelGrouping, continent,

visitNumber, medium). As such, a regression model was used as it is a simple model that estimates the relationship among the variables. The EDA conducted above provided us with some initution of which variables to use, however, our group felt that the model considered too many inputs and it may result in overfitting. Hence, we decided to run LASSO on our training set as it helps to identify and remove some variables.

An important aspect of LASSO's algorithm is that it can provide a more accurate prediction by shrinking and eliminating coefficients that are not fitting, thus, this will assist in reducing variance. Additionally, the cross validation function in LASSO helps us to identify the best overall model. LASSO offers us two model options "lambda.min" - the best performing model and "lambda.1se" - the simplest model one standard deviation away. Our group decided to use "lambda.min" , which consisted of 19 variables (Appendix 1) as it would result in the best performing model. Using the variables identified in "lambda.min", we set out to predict our transactional revenue on the testing data for our first entry on Kaggle.

4.2.1 LASSO Model Evaluation

Our first entry yielded a root mean square error of 1.6410 which ranked us 2928 on the charts. Although this was a decent attempt, running a regression using LASSO has certain flaws. For instance, LASSO cannot identify the p-values for the coefficients, as such, we could not determine which variables were significant. Furthermore, we noticed that was a class imbalance in our dataset as 98.6% of visitors had no transactional revenue. As such, there might not be sufficient identifiable patterns in our minority group (1.4% who had transactional revenue) to appropriately represent its distribution. Our group then pondered over whether adding other additional variables and running LASSO again would yield a better result given the class imbalance present in the dataset. Thus we realised that a regression model might not be the correct approach. As such, we decided to explore a different model that was better equipped to handle the class imbalance.

4.3 Random Forest

The random forest algorithm was chosen as our group's exploratory model because it produces a great result most of the time without advanced tuning of parameters. We have also determined the Google Store case as a combination of a classification

and regression data science project, and this is where Random Forest shines in. Random Forest can be used for both classification and regression, and the good thing is it is also relatively easy to code, even though it runs at a slow pace as more variables are added.

As an exploratory model, one important feature of the algorithm is the ability to determine the important variables so that we can prevent overfitting. Random Forest has ways of showing feature importance: (1) by IncNodePurity which is calculated by looking at a particular feature reduce impurity across all trees in the forest and (2) by IncMSE which shows the increase in mean-squared error of predictions (estimated with out-of-bag cross validation) as a result of variable j being permuted. IncNodePurity is biased and IncMSE, the more robust and informative measure, should almost always be used. However, the limitation is that IncMSE takes a longer time to compute.[7]

The ranger package was used because it optimises for larger dimensions of data, which the traditional Random Forest package in r does not.

4.3.1 Random Forest Model Evaluation

After iterations of finding the best variables based on IncMSE or 'permutation' importance in ranger, we obtained a RMSE on Kaggle of 1.6472. We did not do very well with the random forest, perhaps it is due to a coding error (we realised it was due to wrong removal of duplicates) or wrong choice of variables.

4.4 Insights from initial models

We note that certain variables have higher predictive power, using the Random Forest %IncMSE measure. The top variables that predict transaction revenue are pageviews, subContinent and hits.

This might come as no surprise, as only users who are more interested in the products will have higher pageviews. In addition, users from the subContinent 'North America' have a higher frequency of users purchasing items too.

[7] Soren Havelund Welling. (2015, Jul). In a random forest, is larger %IncMSE better or worse? Cross Validated. Retrieved from https://stats.stackexchange.com/questions/162465/in-a-random-forest-is-larger-incmse-better-or-worse

5. Classification of users into paying and non-paying

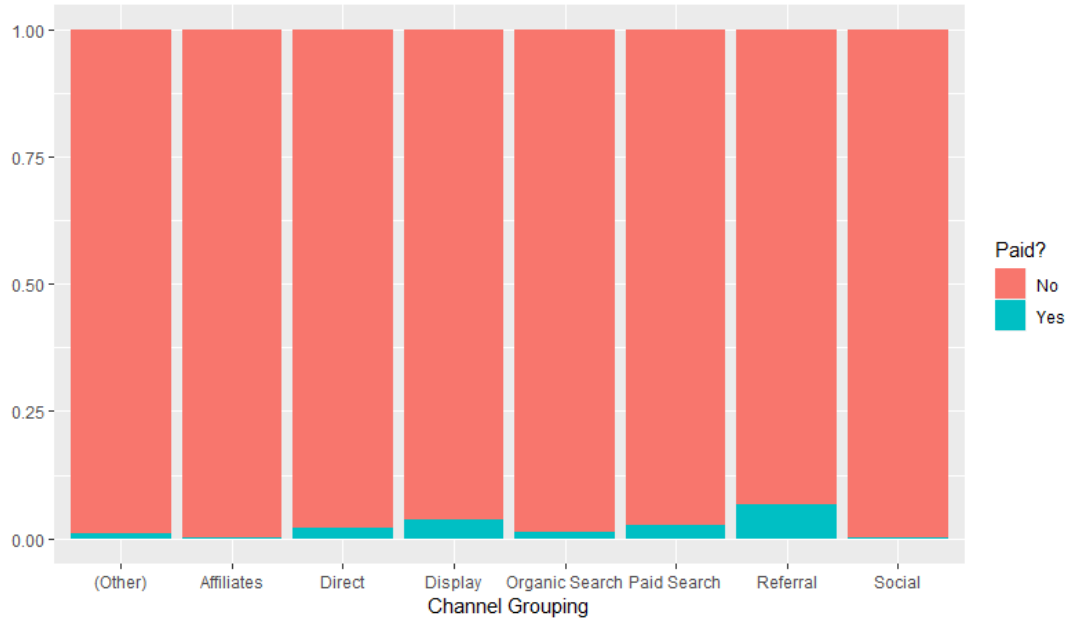
There are only 11,515 visits with revenue as mentioned above, which is only 1.27% of the total observations from the train data. This matches the number of missing values found in transaction revenue previously. Our previous models were attempting to explain variations of these observations with transaction revenue. However, this means that we are only explaining a very small portion of the data. Thus, we will be classifying our observations into two categories: transaction revenue, and no transaction revenue.

As a result, we will first run a model that identifies whether a visitor will make a transaction and then predict the amount of revenue generated from the visitor.

6. Exploratory Data Analysis (Part 2)

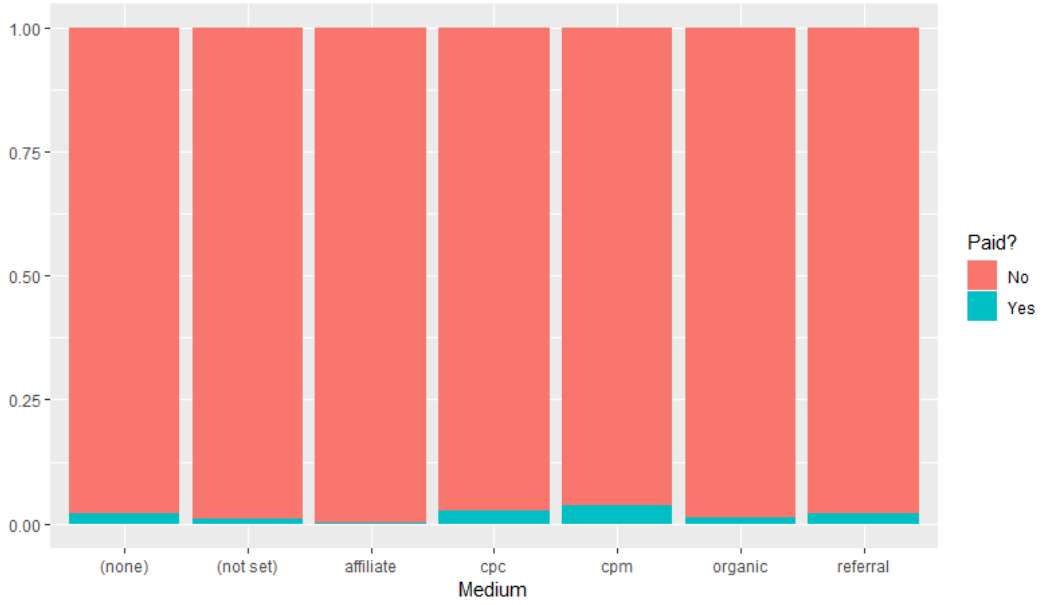
We will analyse the data again after the data has been classified into their categories accordingly. For those who transacted, we assigned a value of 1 and for those visitors who did not transact, we assigned a value of 0.

6.1 Channel Grouping



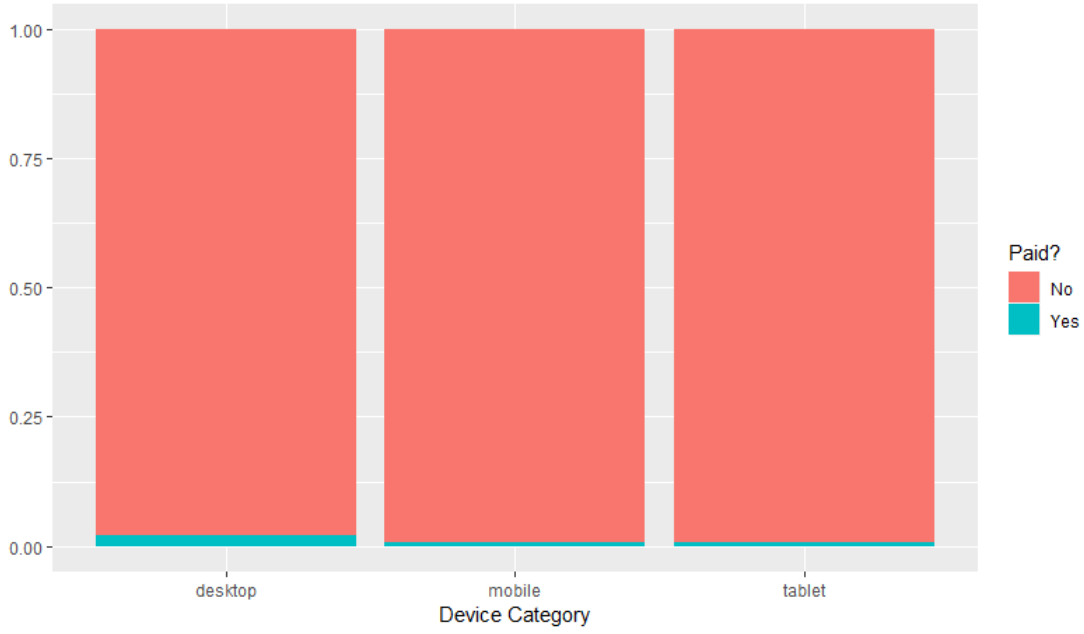
This plot shows that certain channel groupings such as ‘Referral’ and ‘Display’ have a higher proportion of paying customers in their own categories. Perhaps the effectiveness of word-of-mouth marketing is demonstrated through referrals. This is consistent with what we found with our exploratory models and we will keep this variable.

6.2 Medium



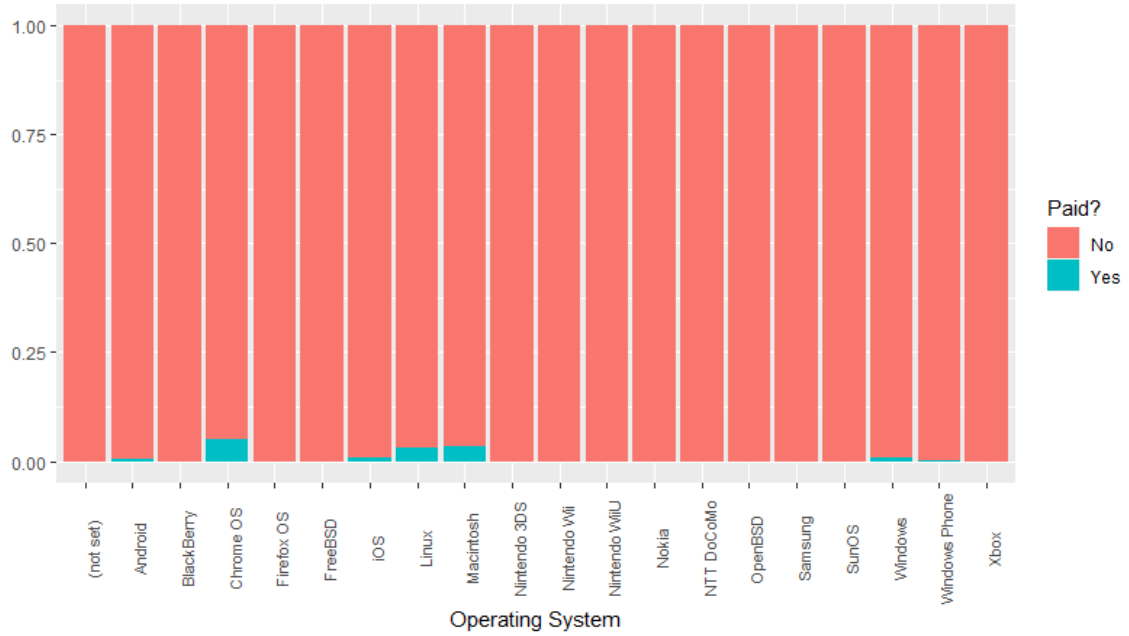
This variable 'medium' might explain some variation due to 'cpm' and 'cpc' variables having slightly higher proportions. We can test this variable's importance later on, or feature engineer it along with another variable.

6.3 Device Category



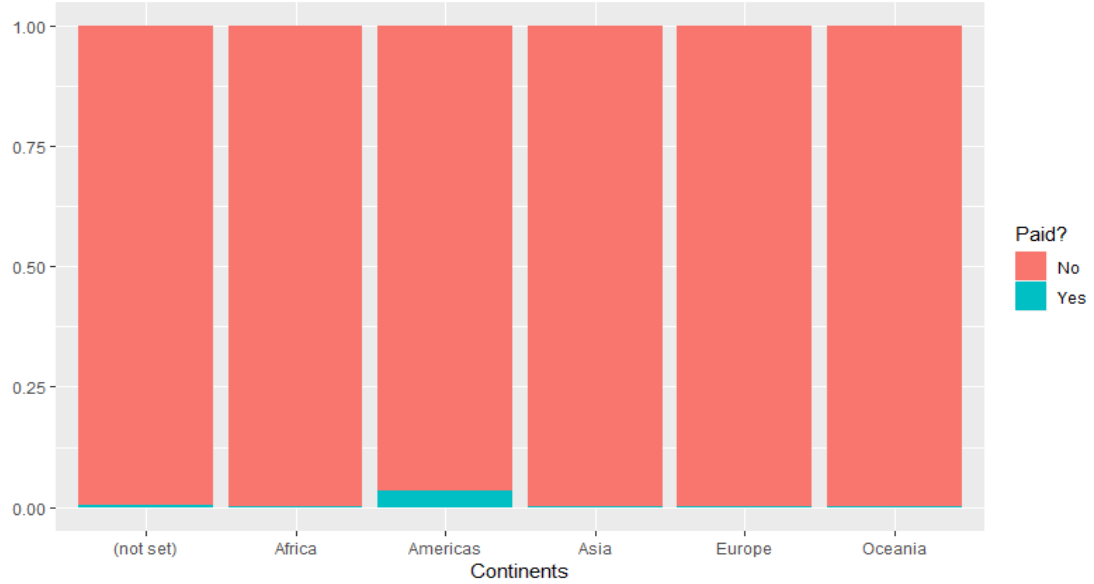
For Device category, the proportion of paying customers is the highest for desktop which is possibly due to desktop being a common means to shop as compared to the other two devices, as mentioned in *3.3.5 Device Categories*. We will feature engineer a binary variable for desktop to include it in the model.

6.4 Operating System



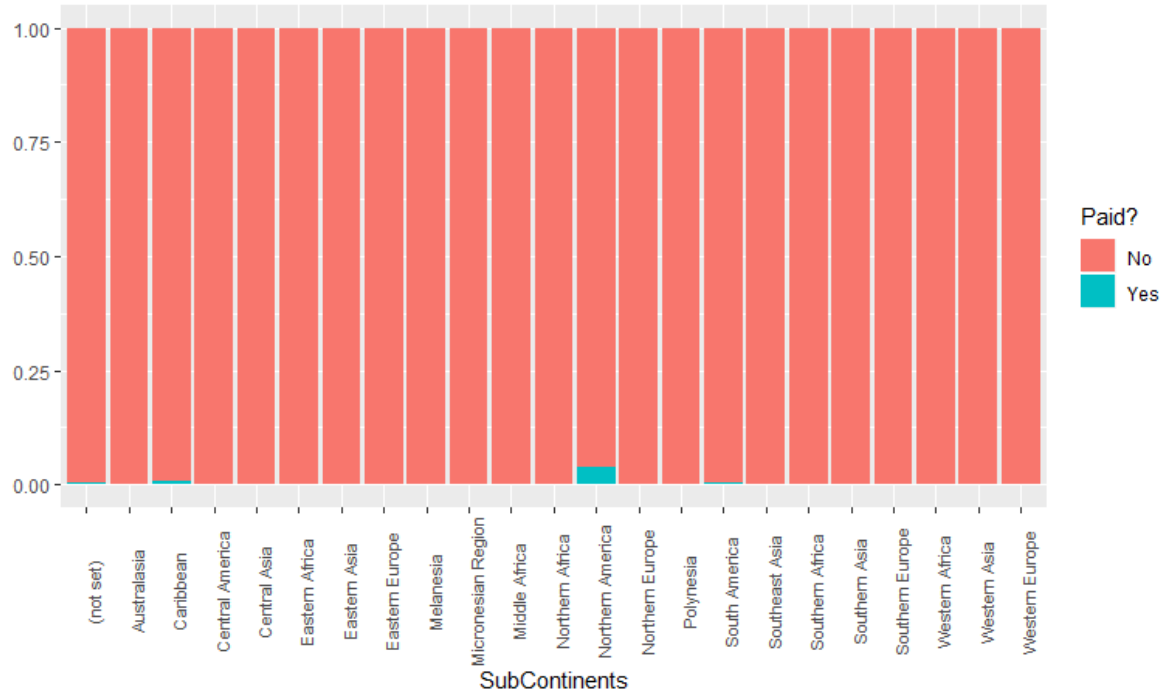
This plot relating to operating system is very informative, as it shows that users on 'Chrome OS', 'Linux' and 'Macintosh' are much more likely to purchase an item. This might be due to certain characteristics of users using these operating systems. For instance, Macintosh users might be more well-off and willing to shop, while Linux and Chrome OS users are likely geeks who are fans of Google. We might be able to feature engineer a binary variable that separates this three variables from the rest.

6.5 Continent



Similar to the results obtained from the Continent plot in our first EDA, Americas shows the highest proportion of paying customers.

6.6 Sub Continents



From this graph, it appears that although Americas have the highest proportion of paying visitors, the proportion is particularly high for visitors from Northern America.

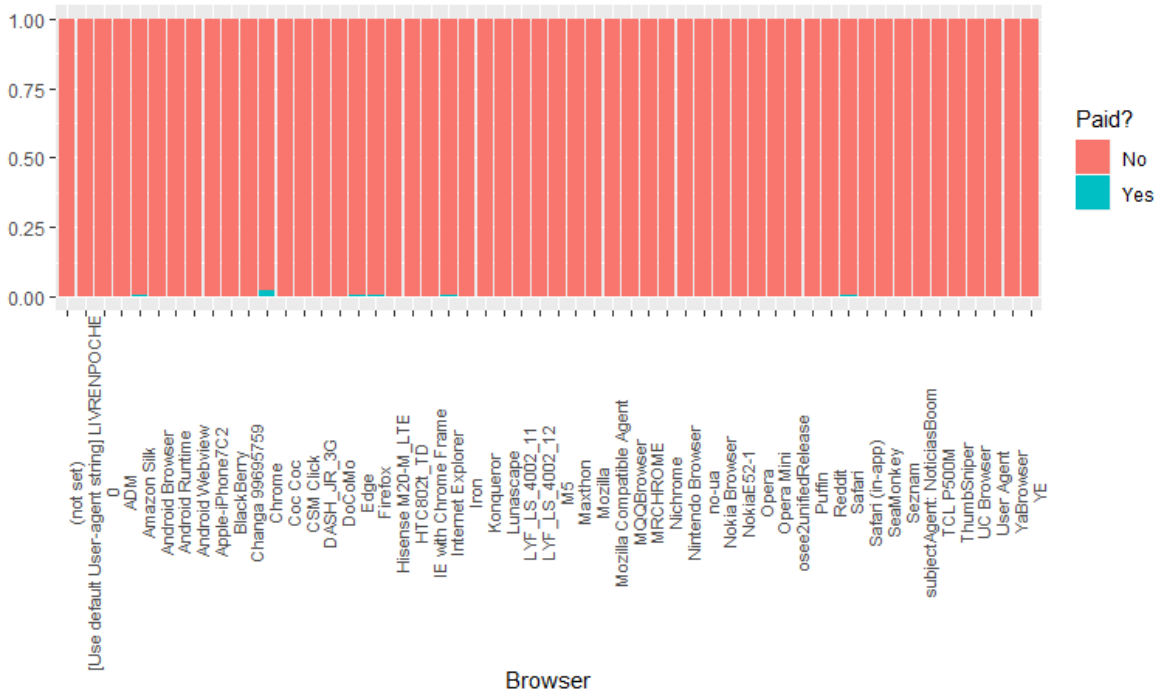
6.7 Country

Due to the above results, we also looked into the proportion of paying visitors for different countries. As there are too many countries included in the dataset, it is impossible to plot a useful graph of the distribution. However, it is noted that United States of America have the highest proportion of visitors. This is likely due to Google being founded in the United States and hence, the greater interest in purchasing Google merchandise.

6.8 Pageviews & hits

In our first EDA, we had expected pageviews and hits to be useful variables in predicting transaction revenue and this was confirmed through the %IncMSE measure of our earlier Random Forest model. We will hence be including them in our revised model.

6.9 Browser



Chrome users have the highest proportion of visitors with transaction revenue. This may be because it is the category with the most users and Chrome is also the browser of Google, which might signify that Google fans are more likely to purchase from GStore.

7. Model Evaluation & Refinement

7.1 Feature engineering

One way to improve on our model is to do some feature engineering. For instance, we can look at the graphs that fills the portion of users who purchased something as '1' and others as '0'. We noted that perhaps browser == Chrome may have some predictive power. Hence, we decided to classify browser == Chrome as '1' and others as '0' and add it into our model.

Another insight we got from EDA is that users with features (deviceCategory == desktop & operatingSystem == "Chrome OS", "Macintosh" or "Linux") tend to buy more than other combinations. As such, we will make a binary variable that separates these combinations from the rest.

7.2 Revised Random Forest Model

After adding the two manually-made variables mentioned above, we found that both variables might explain some of the variation in our results. However, the second variable (the combined one) is found to be more important based on the criteria of IncMSE (or permutation in the ranger package). With this new model, it achieved a RMSE on Kaggle of 1.5616.

7.3 Trial of new model - Gradient Boosting Machines (GBM)

The idea of boosting came from the idea to combine weak models to generate a powerful one. A weak model is one that is slightly better than random performance and through the development of weak learns that describe the remaining difficult observations, we generate an ultimate model that does pretty well. Gradient boosting involves three functions: (1) Loss function, (2) A weak learner to make predictions and (3) An additive model to add weak learners to minimise the loss function.

Decision trees are used as the weak learner in GBM and the additive model adds trees with a gradient descent procedure, reducing the loss.

Under GBM, we then have parameters to help refine the model, namely tree constraints, learning rate, stochastic gradient boosting and penalised gradient boosting.

Both GBM and RF are ensemble learning methods. GBM and RF differ in the ways the trees are constructed: the order and the way results are averaged. It has been found that GBM performs better if the parameters are carefully tuned.

Under our GBM model using the same variables as in RF, we achieved a Kaggle RMSE of 1.5394. Looking at our output, we see negative values for PredictedLogRevenue, which we changed to 0. This gave us the ultimate RMSE of 1.5387.

8. Conclusion

Our final model(GBM) ranked us at 2739 improving our ranking by over 100 positions from our initial rank of 2928. In general, the lower the RMSE, the better. Hence, our final RMSE value of 1.5387 depicted that our model's predictive value was decent but it still had room for improvement.

This project was extremely challenging and tested us intellectually. For many of us, it was the first time that we had to build a model from start to finish. Throughout the process, we learnt that there is no hard and fast rule to build a model. It requires many instances of trial and error by exploring and changing the inputs in our model. Additionally, cleaning and understanding the dataset is vital to build a coherent model. Initially, we did not notice the class imbalance in our dataset (98.6% did not purchase anything) and this led to hiccups along the way while we were trying to refine our initial LASSO model. Thus, only after a second exploratory data analysis was conducted did we realise that due to the class imbalance a different model would be more useful to predict the transactional revenue.

In sum, the entire process from start to finish was an insightful experience that further developed our analytical and problem solving skills. This project developed our ability to learn independently and encouraged us to discover new methods in analytics apart from what was taught in the classroom.

9. References

Brownlee. (2016, Sep). *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning.* Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Donges, N. (2018). *The Random Forest Algorithm.* Machine Learning-Blog. Retrieved from https://machinelearning-blog.com/2018/02/06/the-random-forest-algorithm/

Forbes Communication Council. (2018, Mar). *12 Major Differences Between Mobile And Desktop Marketing.* Forbes. Retrieved from https://www.forbes.com/sites/forbescommunicationscouncil/2018/03/23/12-major-differences-between-mobile-and-desktop-marketing/#76ccd7eb5d6b

Google Analytics. (2018). *Analytics Help - Default channel definitions.* Google. Retrieved from https://support.google.com/analytics/answer/3297892

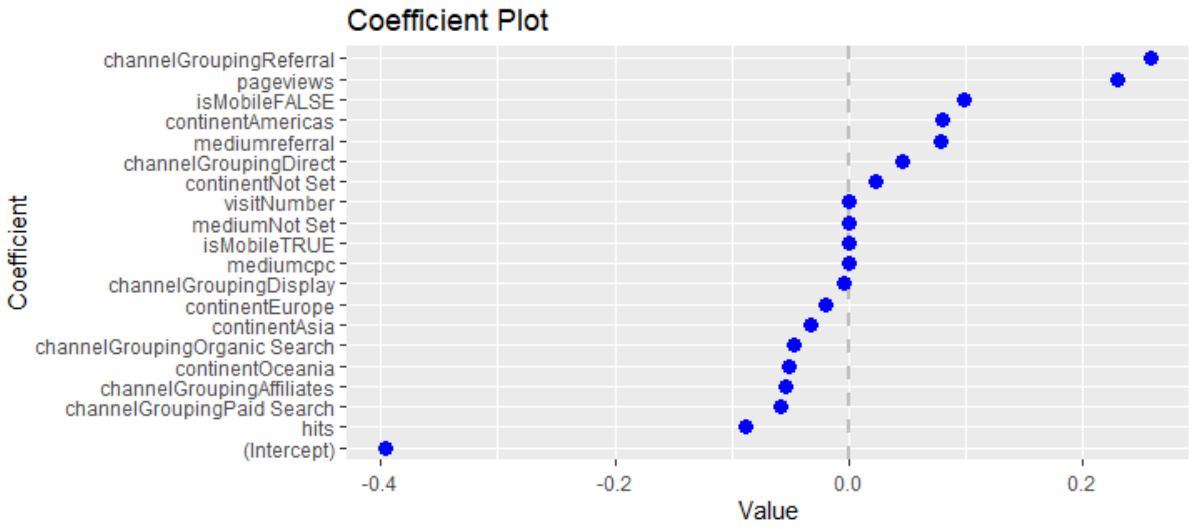
Google Analytics. (2018). *Analytics Help - Traffic Source Dimensions.* Google. Retrieved from https://support.google.com/analytics/answer/1033173?hl=en&ref_topic=6010089

Investopedia. (n.d.) *Cost per Thousand - CPM.* Retrieved from https://www.investopedia.com/terms/c/cpm.asp

Mr Knoot. (2018). *GStore: Crafting Reasonable Models.* Kaggle. Retrieved from https://www.kaggle.com/mrknoot/gstore-crafting-models-manual-features-xgb

Soren Havelund Welling. (2015, Jul). *In a random forest, is larger %IncMSE better or worse?* Cross Validated. Retrieved from https://stats.stackexchange.com/questions/162465/in-a-random-forest-is-larger-incmse-better-or-worse

10. Appendix



Appendix 1 - LASSO Coefficient Plot